CANCER DIAGNOSIS USING DATA MINING TECHNOLOGY

Muhammad Shahbaz[1], Shoaib Faruq[2], Muhammad Shaheen[1], Syed Ather Masood[2]

[1]Department of Computer Science and Engineering, UET, Lahore, Pakistan
Muhammad.Shahbaz@gmail.com, M.Shahbaz@uet.edu.pk
[2] Department of Engineering Management, EME College, NUST, Rawalpindi, Pakistan.
athermasood2000@hotmail.com

Abstract: Cancer is a set of diseases in which some cells of the body grow abnormally. These cells then destroy other surrounding cells and their normal functions. Cancer can spread throughout the human body. Since it is a very treacherous disease its diagnosis is very important. In some forms it spreads within days. So the diagnosis of cancer at early stages is very important. The challenge is to first diagnose the main type and then its subtypes. This research uses data mining classification tools to make a decision support system to identify different types of cancer on the Genes dataset. Data mining technology helps in classifying cancer patients and this technique helps to identify potential cancer patients by simply analyzing the data.
[Muhammad Shahbazl, Shoaib Faruq, Muhammad Shaheen, Syed Ather Masood. **Cancer Diagnosis Using Data Mining Technology.** Life Science Journal. 2012;9(1):308-313] (ISSN:1097-8135). http://www.lifesciencesite.com. 44

Keywords: Data Mining; K-Nearest Neighbors; Naïve Bayesian; SVM, Classification; Cancer

INTRODUCTION

Cancer is normally diagnosed by examining the cells using a microscope. Imaging tests like computerized tomography (CT) or mammography help in indicating the possible presence of cancer by depicting an abnormal growth or mass. Final decision is usually taken by having different kinds of lab tests of the patient and observing closely the cancer cells under study.

Another method used by Doctors is called biopsy. Biopsy is done by surgery. Doctors take a sample of the tissue that is understudy. This sample is then examined with the help of a microscope. The appearance of normal cells is uniform; they are organized in order and are of equal size. Cancer cells are different than normal cells. They are in dispersed order, their sizes are different and they are not structured well.

The problem with this is that a medical image such as CT scan or MRI cannot show all the patterns and information for a particular type of cancer or subtypes of cancer. Another issue is that a doctor with his/her naked eye and a microscope cannot remember a large number of patterns of the disease.

It is frightening for a patient to know that he/she has cancer. A patient can lose all hope after being diagnosed with cancer. Therefore cancer diagnosis is a process that needs proper care and patience on both sides i.e. the patient and doctor/hospital.

Early diagnosis of cancer can help save the life of a patient because Cancer cells cause destruction to other cells and spread to other parts of body very quickly. If it is diagnosed in the early stage, the treatment begins earlier and this can prevent further spread of the disease.

Existing diagnosis system at cancer hospitals: Currently cancer diagnosis system in hospitals is manual. For example when a patient is registered he/she has to go through radiology test process i.e. X-rays, CT or MRI. Radiologist gives his remarks on the test report. After this process an expert doctor reviews the X-rays/CT/MRI and gives his remarks. In some types of cancer the diagnosis is based on the final decision by the doctors e.g. breast and lung cancer, but in other types of cancer like carcinoma some other tests are also required like biopsy. In a manual system the radiologist and the doctor diagnose cancer. This process is slow as after the radiologist's review the doctor has to review also and give his/her remarks and finally tell if the cancer is present or not. The need is to automate this process to make the cancer diagnosis efficient and fast with the use of state of the art technology.

Genes and their importance in Cancer Diagnosis: Genes provide very valuable information which can be used to study any disease in depth. Study of genes from a cancer patient helps us diagnose cancer and differentiate between types of cancer. It also helps in separating the healthy people from the patients. Genes contains infinite patterns that cannot be recorded manually using a microscope. DNA Micro Arrays are used to study the information obtained from Genes.

DNA Micro Arrays: DNA microarrays are the latest form of biotechnology. These allow the measurement of genes expression values simultaneously from hundreds of genes. Some of the application areas of DNA microarrays are obtaining the genes values

from yeast in various ecological conditions and studying the gene expression values in cancer patients for different cancer types. DNA Microarrays have huge potential scientifically as they can be useful in the study of genes interactions and genes regulations. Other application areas of DNA microarrays are clinical research and pharmaceutical industry [1].

Data Retrieval from DNA Micro Arrays: Gene expression data is retrieved from DNA microarray through Image processing techniques. Data for a single gene consists of two intensity values of fluorescence i.e. Red and Green. These intensities represent expression level of gene in Red and Green labeled mRNA samples. Image of a microarray is scanned. This image is then processed through image processing techniques [1].

Image Processing: DNA microarrays are scanned using laser scanners and its output is stored as 16-bit image. Image format is in DICOM. As DICOM is a standard for storing medical images. This image is considered raw input. In order to measure the accurate transcript wealth, different image processing methods are employed [1].

The steps for processing the scanned image from a DNA Micro Array are as follows.

Automatic Address: To get accurate values of intensities from microarray data we need to identify the address/location of each gene point or spot. This is known as automatic addressing and it is used to assign the spot coordinates. Accurate identification of the locations of the spots is mandatory to calculate the spot intensities.

Segmentation: Segmentation is a technique which separates the point of interest from the background. It is used to get the actual values of gene spots and differentiate from background of the image.

Intensity Extraction: Intensity extraction is an important step in image processing. Measurement of the Intensities of spots, background and quality measurements are done in this step.

Signal: The sum of pixel intensities within a particular spot is called signal. The collective amount of cDNA hybridized at the marked DNA sequence is represented by this sum.

PROBLEM STATEMENT AND RELATED DATA

Sample data under study is gene expression data of cancer type leukemia and it is freely available for download at [15]. The dataset consists of 72 bone marrow samples. Samples are from the acute leukemia patients. These samples are from the patients having two types of acute leukemia i.e. acute lymphocytic leukemia (ALL) and acute myelogenous leukemia (AML). First 38 samples are training samples from which first 27(From 1 to 27) cases are ALL and 11 (From 28 to 38) are AML. In these 38 samples 8 out of 27 are T-cell samples and 19 are B-cell samples. [16] The remaining 34 are test samples in which 20 are ALL and 14 AML. In ALL sample 19 are B-cell samples and 1 is T-cell sample. [17]

Each sample contains 7129 human genes expressions spotted on a DNA microarray as described above.

Data was in the form of a data file (.dat) and it was converted to comma separated values (.csv) file format using MATLAB. It was then used for further analysis using Data Mining Tool called Rapid miner. Every record had its class attribute. Class attribute was in numeric form. There were three classes

1. AML
2. ALL – B Cell
3. ALL – T Cell

In original dataset ALL – B Cell class was represented by value 0, ALL – T Cell class was represented by 1 and AML was represented by a value 2. For analysis I have changed the class attribute from numeric to character value as follows:

Table 1: Class Label Transformation

Class Attribute	Old Value	New Value
ALL – B Cell	0	ALL-B
ALL – T Cell	1	ALL-T
AML	2	AML

Figure 1 below shows the view of data, first column of data is class. From column 2 to 7130 are gene expression values for each sample of each DNA.

The challenge here is to find out the best classification method that help in identifying the classes present in data.

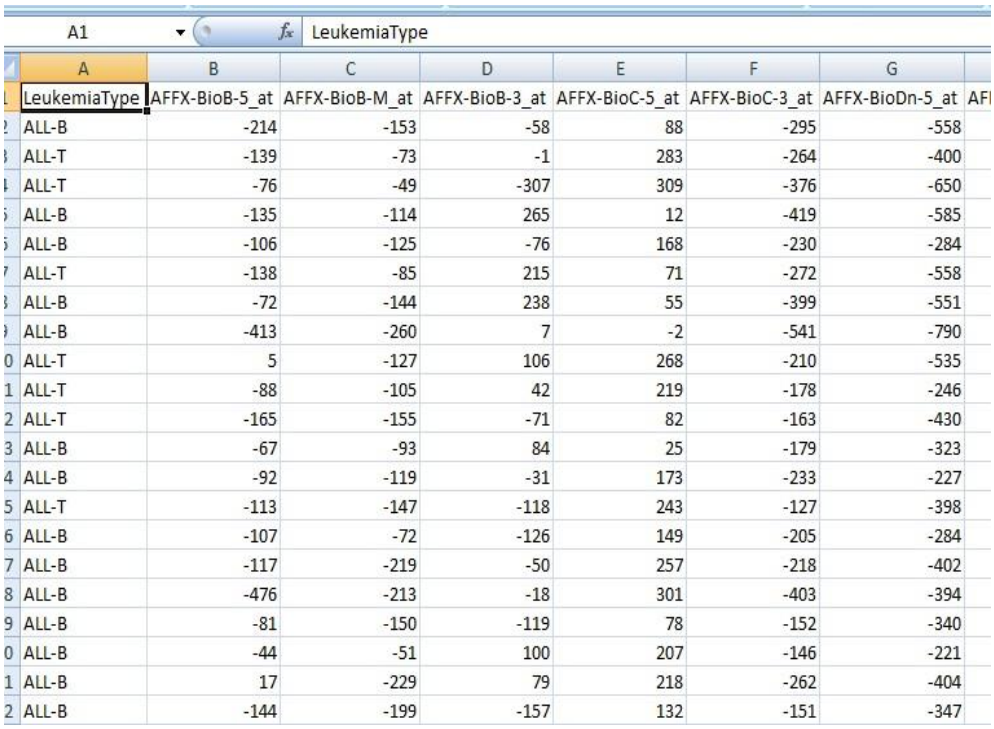
RESULTS AND DISCUSSIONS

Below is a summary of results and performance comparison of the experiments performed above. We have performed experiments using three algorithms Naïve Bayesian, K Nearest Neighbors and SVM. For each result a confusion matrix is presented which shows the actual samples in a particular class and the predicted class. Accuracy of the classification algorithm is also given with the results.

Results for Naïve Bayesian Algorithm:

Table 2: Confusion Matrix for Naïve Bayesian Algorithm

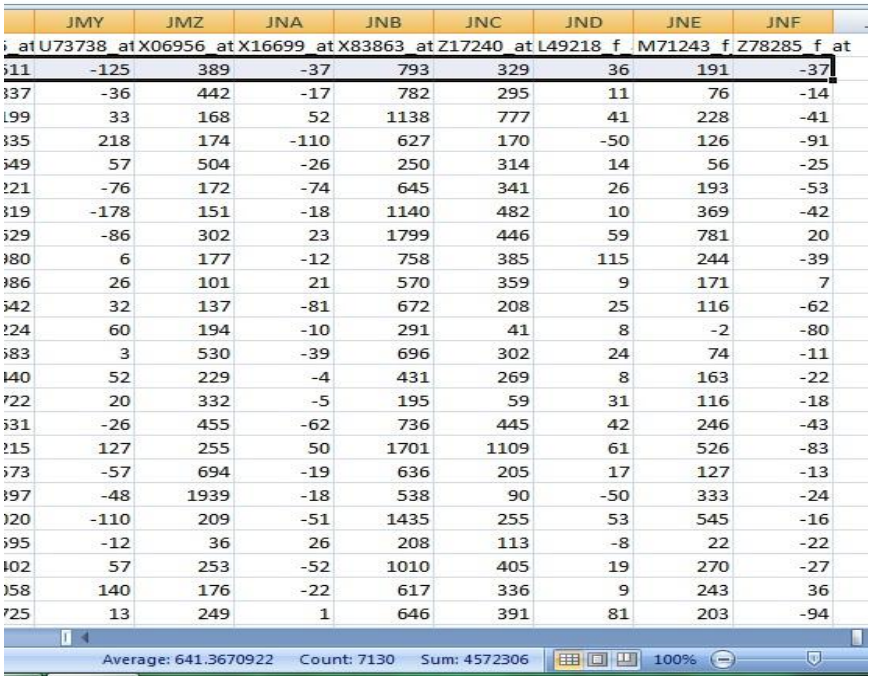
	Actual ALL-B	Actual ALL-T	Actual AML	Class Precision
Predicted ALL-B	37	4	0	90.24 %
Predicted ALL-T	0	5	0	100 %
Predicted AML	1	0	25	96.15 %



A1 LeukemiaType

	A	B	C	D	E	F	G	
1	LeukemiaType	AFFX-BioB-5_at	AFFX-BioB-M_at	AFFX-BioB-3_at	AFFX-BioC-5_at	AFFX-BioC-3_at	AFFX-BioDn-5_at	AFI
2	ALL-B	-214	-153	-58	88	-295	-558	
3	ALL-T	-139	-73	-1	283	-264	-400	
4	ALL-T	-76	-49	-307	309	-376	-650	
5	ALL-B	-135	-114	265	12	-419	-585	
6	ALL-B	-106	-125	-76	168	-230	-284	
7	ALL-T	-138	-85	215	71	-272	-558	
8	ALL-B	-72	-144	238	55	-399	-551	
9	ALL-B	-413	-260	7	-2	-541	-790	
10	ALL-T	5	-127	106	268	-210	-535	
11	ALL-T	-88	-105	42	219	-178	-246	
12	ALL-T	-165	-155	-71	82	-163	-430	
13	ALL-B	-67	-93	84	25	-179	-323	
14	ALL-B	-92	-119	-31	173	-233	-227	
15	ALL-T	-113	-147	-118	243	-127	-398	
16	ALL-B	-107	-72	-126	149	-205	-284	
17	ALL-B	-117	-219	-50	257	-218	-402	
18	ALL-B	-476	-213	-18	301	-403	-394	
19	ALL-B	-81	-150	-119	78	-152	-340	
20	ALL-B	-44	-51	100	207	-146	-221	
21	ALL-B	17	-229	79	218	-262	-404	
22	ALL-B	-144	-199	-157	132	-151	-347	

Figure 1: Dataset showing class and attribute values



	JMY	JMZ	JNA	JNB	JNC	JND	JNE	JNF
_at	U73738_at	X06956_at	X16699_at	X83863_at	Z17240_at	L49218_f_	M71243_f	Z78285_f_at
511	-125	389	-37	793	329	36	191	-37
337	-36	442	-17	782	295	11	76	-14
199	33	168	52	1138	777	41	228	-41
335	218	174	-110	627	170	-50	126	-91
549	57	504	-26	250	314	14	56	-25
221	-76	172	-74	645	341	26	193	-53
319	-178	151	-18	1140	482	10	369	-42
529	-86	302	23	1799	446	59	781	20
980	6	177	-12	758	385	115	244	-39
986	26	101	21	570	359	9	171	7
542	32	137	-81	672	208	25	116	-62
224	60	194	-10	291	41	8	-2	-80
583	3	530	-39	696	302	24	74	-11
440	52	229	-4	431	269	8	163	-22
722	20	332	-5	195	59	31	116	-18
531	-26	455	-62	736	445	42	246	-43
215	127	255	50	1701	1109	61	526	-83
573	-57	694	-19	636	205	17	127	-13
397	-48	1939	-18	538	90	-50	333	-24
020	-110	209	-51	1435	255	53	545	-16
595	-12	36	26	208	113	-8	22	-22
402	57	253	-52	1010	405	19	270	-27
058	140	176	-22	617	336	9	243	36
725	13	249	1	646	391	81	203	-94

Average: 641.3670922 Count: 7130 Sum: 4572306 100%

Figure 2: Showing the last columns of data

As described earlier leukemia data set has three classes ALL-B, ALL-T and AML. First column in the matrix represents the predicted class and subsequent columns represent the actual number of occurrences in each class. If we look at column 2 of row 2 it shows the value 37. It means that there were actually 37 samples of class ALL-B in our data and they are classified by Naïve Bayesian algorithm

correctly. Now we move on to row 2 and column 3 it shows a value 4. It means that these 4 samples are incorrectly classified. In last column of row 2 there is a 0 value which means that there are no samples which are incorrectly classified as AML. To see the correctly classified instances we must see values in diagonal i.e. 37, 5 and 25. It means that there are 5 samples incorrectly classified represented by row 2 column 3 with value 4 and row 4 column 2 with value 1. We can calculate the accuracy of the algorithm by simple method that there are total 72 samples and 5 fall to an incorrect class so the accuracy of classification is 95% approximately.

Last column in confusion matrix shows the precision of each predicted class. Precision for class ALL-B is 90.24 %, for ALL-T is 100 % and for AML it is 96.15% for Naïve Bayesian classifier.

Results for K Nearest Neighbor Algorithm:

Table 3: Confusion Matrix for K Nearest Neighbor Algorithm

	Actual ALL-B	Actual ALL-T	Actual AML	Class Precision
Predicted ALL-B	38	2	5	84.44 %
Predicted ALL-T	0	7	0	100 %
Predicted AML	0	0	20	100 %

As described previously leukemia data set have three classes ALL-B, ALL-T and AML. First column in the matrix represents the predicted class and subsequent columns represent the actual number of occurrences in each class. According to K Nearest neighbor algorithm there are 38 instances in row 2 and column 2 that are classified correctly. In next column of row two the value is 2 that depicts that there are 2 samples which are incorrectly classified as ALL-T. In column 3 of row 2 value of 5 shows that there are 5 samples incorrectly falling in AML category/class. For ALL-T if we look at diagonal values that is row 3 and column 3, 7 instances are correctly classified by KNN algorithm. There are 20 instances which are correctly classified as AML in last row and last column. Accuracy for this algorithm can be calculated by subtracting the incorrectly classified instances. We have 72 samples, out of which 7 are incorrectly classified by KNN so we have accuracy value 90.72% for this method.

Class precision is shown in the last column of the confusion matrix for predicted classes. Precision for class ALL-B is 84.44 %, for ALL-T is 100 % and for AML it is 100% for K Nearest Neighbor Classifier.

Results for SVM Algorithm:

Table 4: Confusion Matrix for SVM learning algorithm

	Actual ALL-B	Actual ALL-T	Actual AML	Class Precision
Predicted ALL-B	38	4	3	84.44 %
Predicted ALL-T	0	5	0	100 %
Predicted AML	0	0	22	100 %

For SVM Classification method we have one value in confusion matrix at row 2 and column 3 which is 4 this is incorrectly classified by SVM. Another value incorrectly classified is at row 2 column 4 it has value 3. There are 7 samples which are incorrectly classified by SVM. Accuracy for this method is 90.27%.

Class precision is shown in the last column of the confusion matrix for predicted classes. Precision for class ALL-B is 84.44 %, for ALL-T is 100 % and for AML it is 100% for SVM classification algorithm.

CONCLUSION

According to results above Naïve Bayesian Classification has the most accurate prediction for leukemia dataset samples. Naïve Bayesian classified 95% of the samples correctly in their respective classes. It has only error rate of 5%. Naïve Bayesian is the best method for classifying DNA Microarray genes expression data.

Figure 3 represents the comparison of the three algorithms and their correctly classified samples / instances. Similarly Figure 4 depicts the number of incorrectly classified instances.

Accuracy for different algorithms is shown in figure 5 i.e. Naïve Bayesian, K Nearest Neighbors and SVM.

SUGGESTIONS AND FUTURE WORK

The results above can be improved by reducing the number of attributes we have in the dataset. This can be done using dimensionality reduction techniques like principle component analysis. Problem with principle component analysis is that we do not have the track of data which is considered redundant by this method. If we can somehow obtain the pattern of data that is redundant and get information about which attributes or values are retained then it would be a great improvement in the classification results of any of the learning algorithm.

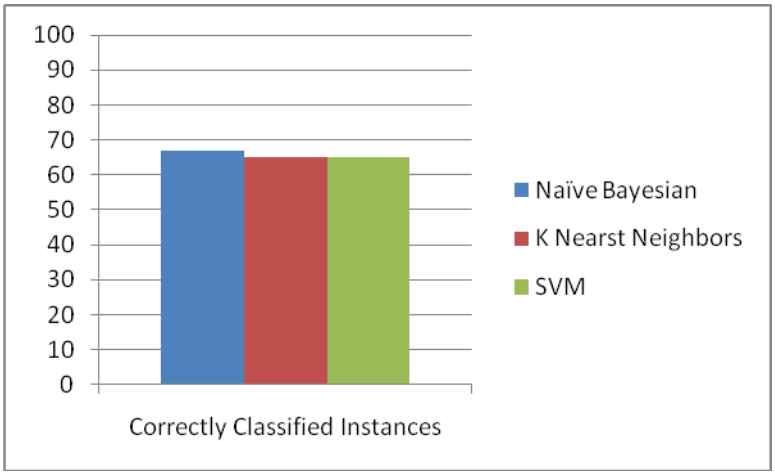


Figure 3: Correctly classified instance by each learning algorithm

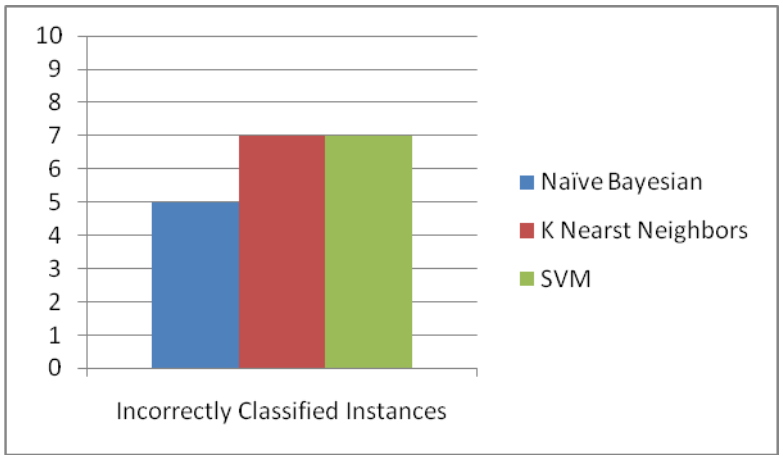


Figure 4: Incorrectly classified instance by each learning algorithm

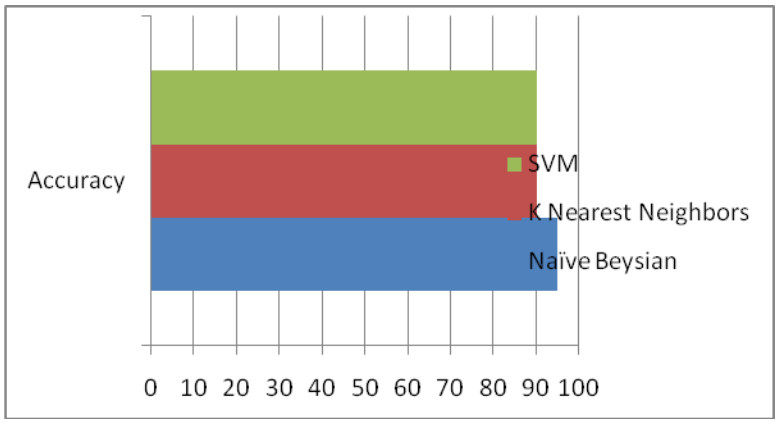


Figure 5: Accuracy for each learning algorithm

REFERENCES

[1]. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.
[2]. http://www.cancer.gov/cancertopics/what-is-cancer.
[3]. Cancer Research UK (January 2007). UK cancer incidence statistics by age. Retrieved on 2007-06-25.
[4]. WHO (February 2006). Cancer. World Health Organization. Retrieved on 2007-06-25.
[5]. American Cancer Society (December 2007). Report sees 7.6 million global 2007 cancer deaths. Reuters. Retrieved on 2007-12-17.
[6]. Duerinckx AJ, Pisa EJ. Filmless Picture Archiving and Communication System (PACS) in Diagnostic Radiology. Proc SPIE 1982;318;9-18. Reprinted in IEEE Computer Society Proceedings of PACS'82, order No 388.
[7]. Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5.
[8]. Principles of Data Mining. Max Bramer, BSc, PhD, CEng, FBCS, FIEE, FRSA, Digital Professor of Information Technology, University of Portsmouth, UK. ISBN-10: 1-84628-765-0.
[9]. Quinlan, J. R., Induction of Decision Trees. Machine Learning, 1986. 1(1): pp. 81-106.
[10]. Tutorial: Introduction to Belief Networks, Teknomo, Kardi. K-Nearest Neighbors Tutorial. http:\\people.revoledu.com\kardi\ tutorial\KNN\
[11]. A Tutorial on Support Vector Machines for Pattern Recognition, CHRISTOPHER J.C. BURGES, Bell Laboratories, Lucent Technologies Gene Expression Profiling based Multi-Class Cancer Classification using AdaBoost and Artificial Neural Network, Gwangju Institute of Science and Technology (GIST)
[12]. Gwangju, Republic of Korea. Classification and diagnostic prediction of cancers using gene, expression profiling and artificial neural networks, JAVED KHAN, JUN S. WEI, MARKUS RINGNÉR, LAO H. SAAL, MARC LADANY, FRANK WESTERMANN, FRANK BERTHOLD, MANFRED SCHWAB, CRISTINA R. ANTONESCU, CARSTEN PETERSON & PAUL S. MELTZER. Nature Publishing Group http://medicine.nature.com

1/19/2012